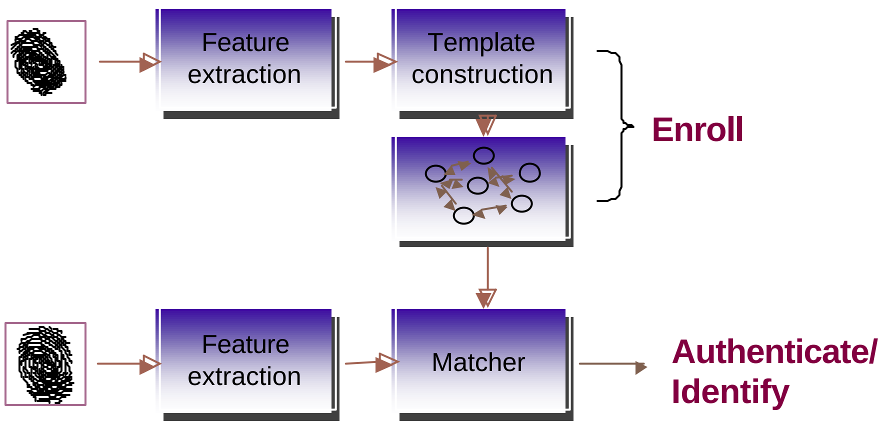
Accuracy Estimation of Biometrics Systems: *The Subsets Bootstrap*

R.M. Bolle, N.K. Ratha, and S. Pankanti
Exploratory Computer Vision Group
IBM Thomas J. Watson Research Center
Yorktown Heights, NY 10598
{bolle, ratha, sharat}@us.ibm.com

Outline

- Biometrics systems
- Performance accuracy
- Probability densities and distributions
- The match score distribution
- Confidence interval
- The Bootstrap
- The Subsets Bootstrap
- Experiments
- Conclusions

Biometrics Systems



Pattern Recognition System

Two hypotheses:

$H_o \quad : \quad b = b',$ the claimed identity is correct

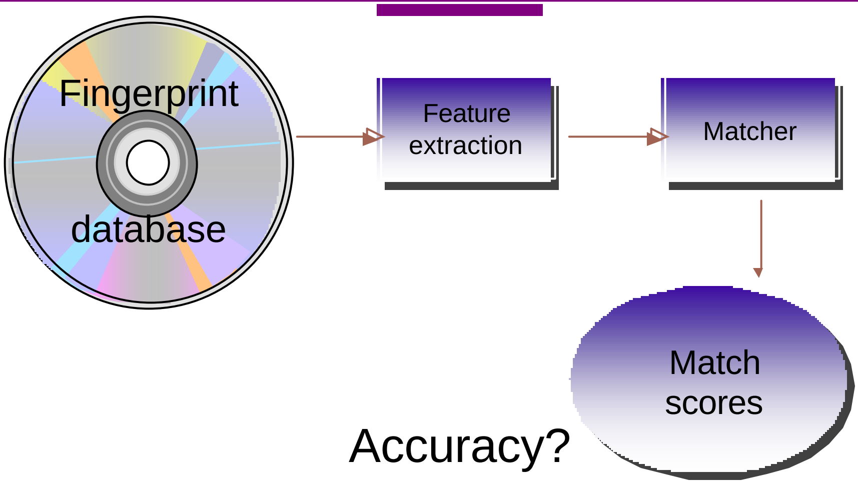
$H_1 \quad : \quad b \neq b',$ the claimed identity is *not* correct

Two errors:

A False Reject: *Decide* H_1 *while* H_o *is true*

A False Accept: *Decide* H_o *while* H_1 *is true*

Performance Accuracy



What is Accuracy?

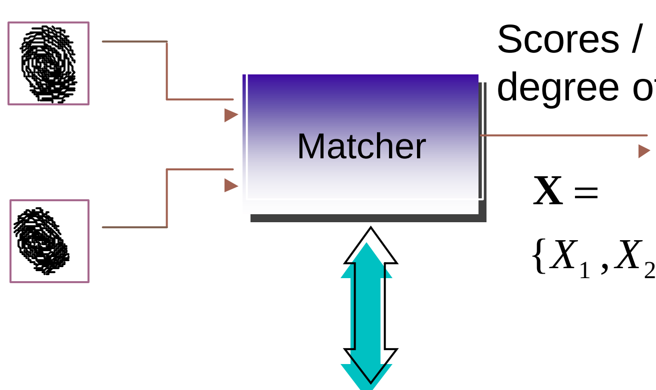
- False reject and false accept rates?
- Equal error rate?
- d-prime?
- Receiver Operator Characteristic curve (ROC)?

- All these are estimates....
- What is the significance of the estimator(s), i.e., confidence intervals?

Confidence Interval Estimation

- **Parametric**
  - **Assumes some parametric form of the underlying distribution**
  - **Binomial distribution of match scores; uses law of large numbers**
- **Non-parametric**
  - **Bootstrap**
- **Both methods assume i.i.d. samples**
  - **independently distributed**
  - **identically distributed**

What is a Matcher?



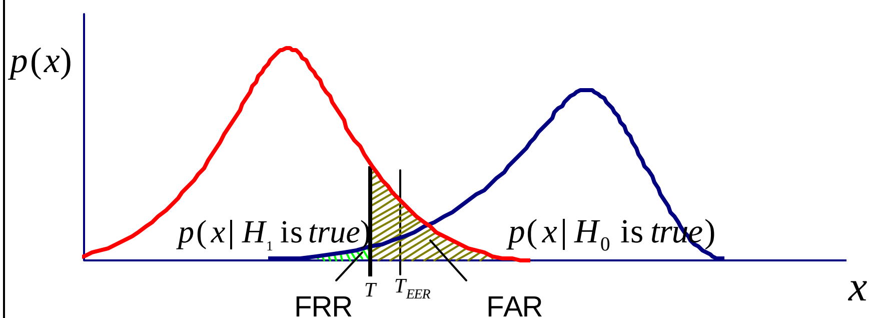
Scores / samples X
degree of similarity

$$\mathbf{X} = \{X_1, X_2, X_3, \ldots\ldots \}$$

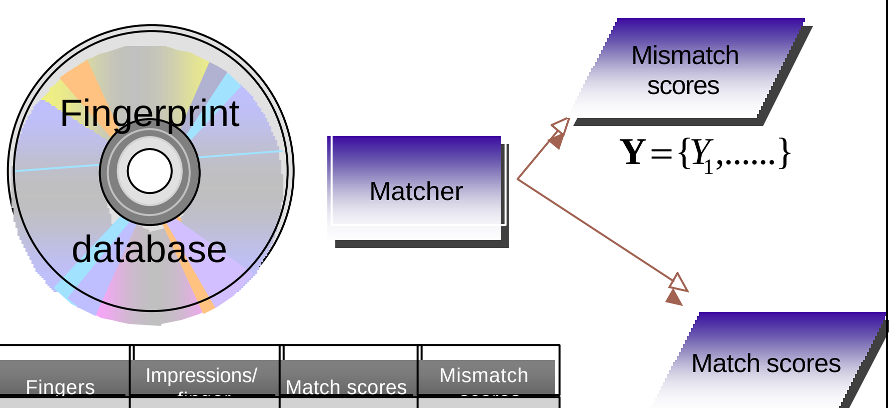
$$p(x); \quad F(x) = \int p(y)dy$$

FARs and FRRs

- **Decide a match if:** $x \geq T$
- **Tradeoff between FRR and FAR**



Generating Match Scores



$$\mathbf{Y} = \{Y_1, \ldots\ldots\}$$

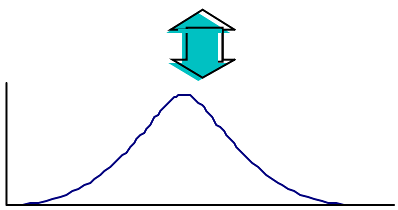
Fingers	Impressions/finger	Match scores	Mismatch scores
D	d	$\frac{Dd(d-1)}{2}$	$\frac{d^2 D(D-1)}{2}$

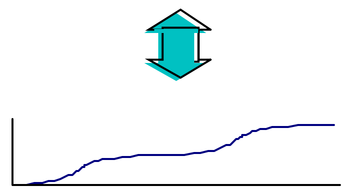
$$\mathbf{X} = \{X_1, X_2, \ldots\ldots\ldots, X_M\}$$

Probability Density

Of match scores $\mathbf{X} = \{X_1, X_2, \ldots\ldots\ldots\}$

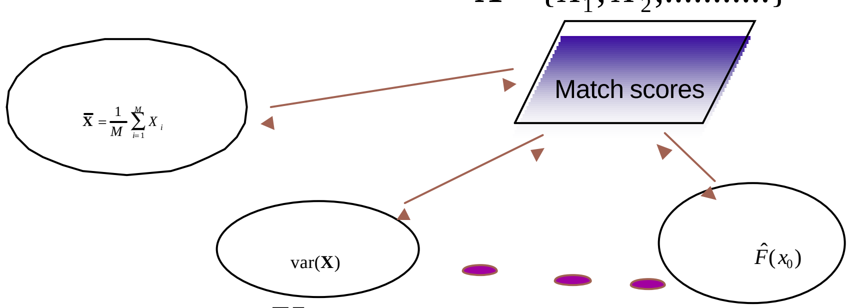
$$p(x \mid H_0 = true) \qquad F(x) = \int_{\infty}^{x} p(y \mid H_0 = true)\, \mathrm{d}y$$





Computing Statistics

$$\mathbf{X} = \{X_1, X_2, \ldots\ldots\ldots\}$$



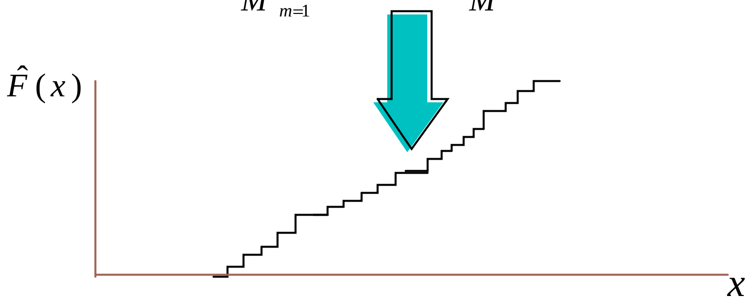
Any statistic of $\mathbf{X}$ is only an estimate:

error measures are statistics, hence,
random variables

Empirical Distribution

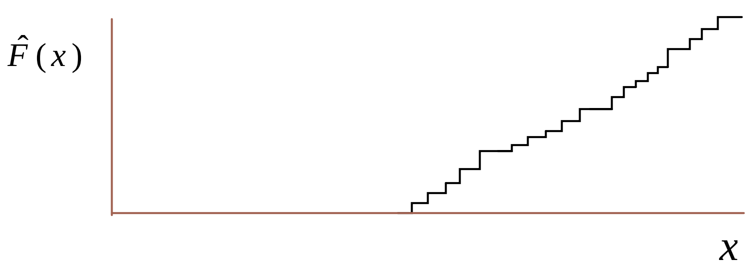
Match scores: $\mathbf{X} = \{X_1, X_2, \ldots, X_M\}$

$$\hat{F}(x) = \frac{1}{M}\sum_{m=1}^{M}\mathbf{1}(X_m \le x) = \frac{1}{M}\#(X_m \le x)$$



The Bootstrap

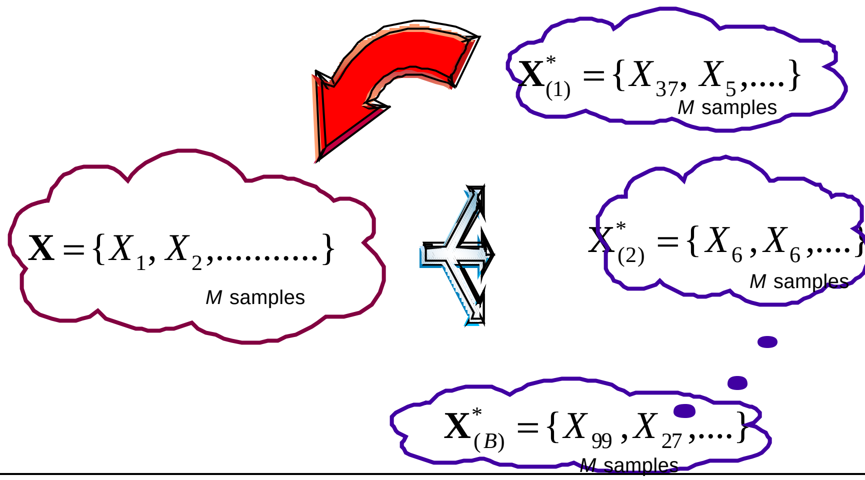
Assume $\hat{F}(x)$ is the true distribution $F(x)$



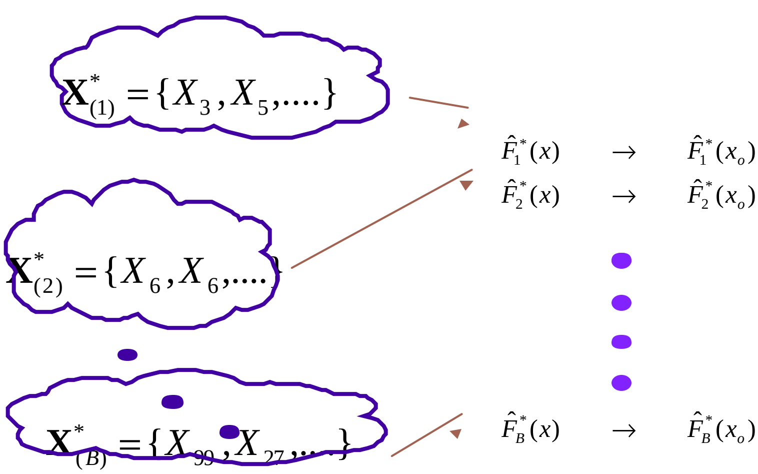
<u>Sample</u> $\mathbf{X}$ **many** **(*B*)** times, with replacement

$$\mathbf{X}^*_{(1)}, \mathbf{X}^*_{(1)}, \ldots, \mathbf{X}^*_{(B)}$$

Sampling with replacement



Bootstrap Estimates



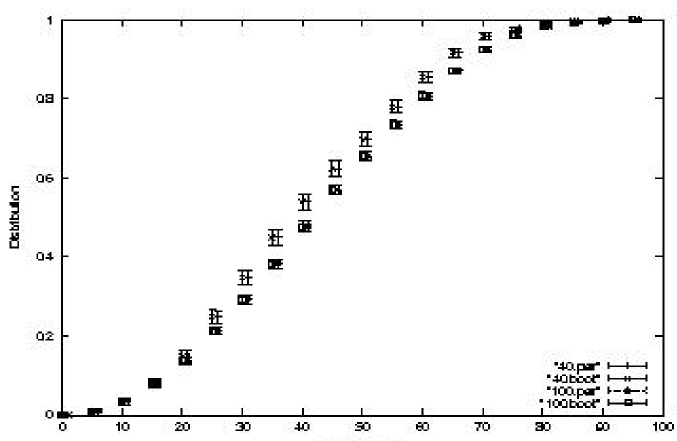
Ordering and Counting

$$\{\hat{F}_1^*(x_o),\quad \hat{F}_2^*(x_o),\quad . \quad . \quad . \quad , \quad \hat{F}_B^*(x_o)\}$$

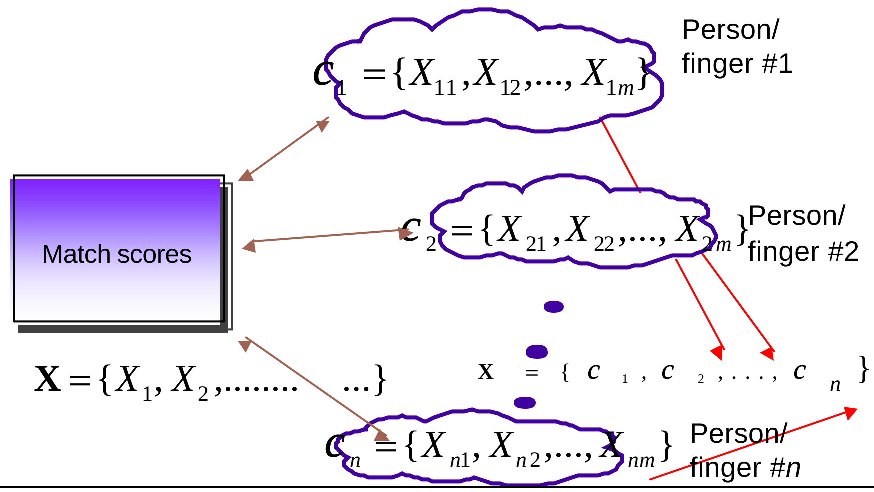
$$(\hat{F}_{(1)}^*(x_o) \;\leq\; \hat{F}_2^*(x_o) \;\leq\; . \quad . \quad . \;\leq\; \hat{F}_{(B)}^*(x_o))$$

Confidence Intervals

False Reject Curve

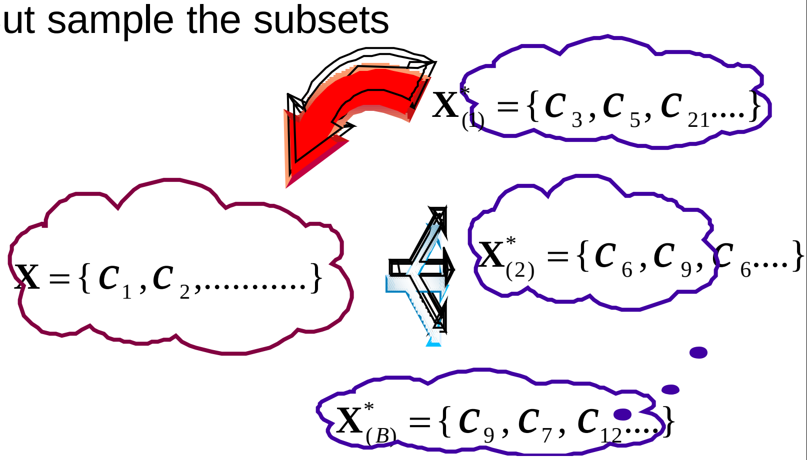


Match Scores not Independent

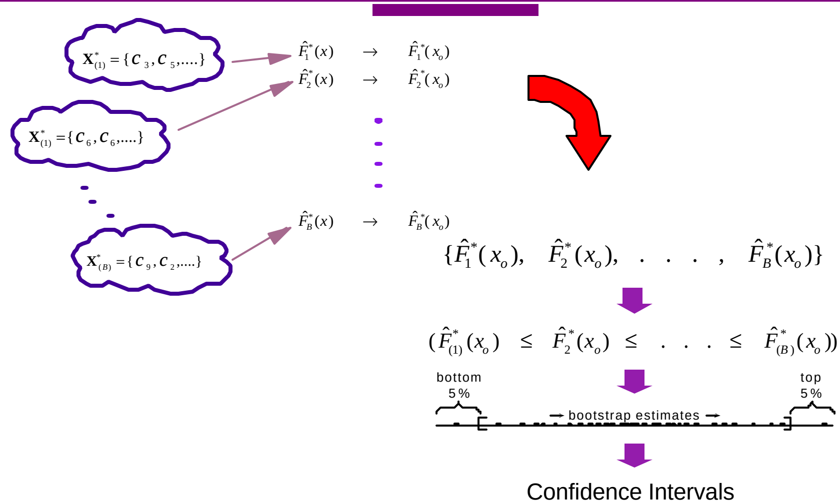


Sampling with Replacement

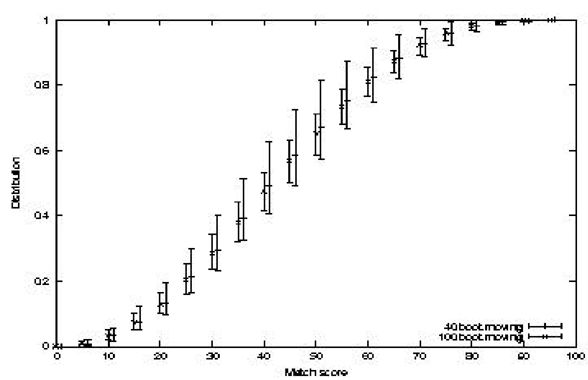
But sample the subsets



Estimates & Confidence Intervals



False Reject Rate Curve



Replication of Dependence

$$\mathbf{X} = \{X_1, X_2, X_3, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{11}, X_{12}, X_{13}, X_{14}, X_{15}, \ldots\ldots, X_M\}$$

$$\mathbf{X}^* = \{X_4, X_{11}, X_2, X_{15}, X_1, X_{12}, X_{11}, X_7, X_2, X_4, X_3, X_{13}, X_9, X_1, \ldots\ldots, X_3\}$$

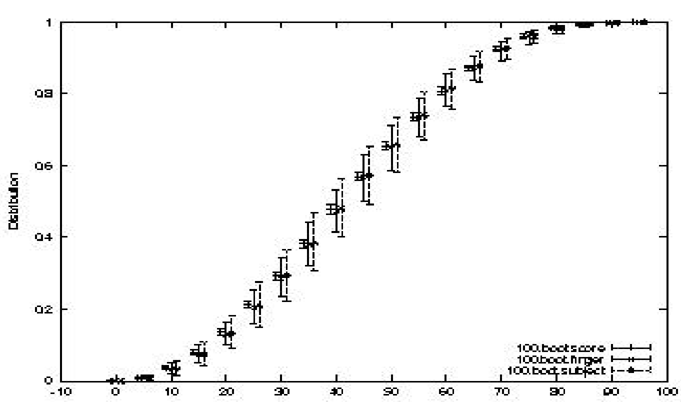
$$\mathbf{X} = \{X_1^a, X_1^b, X_1^c, X_1^d, X_2^a, X_2^b, X_2^c, X_2^d, X_3^a, X_3^b, X_3^c, X_3^d, \ldots, X_m^a, X_m^b, X_m^c, X_m^d\}$$

$$\mathbf{X} = \{ \ c_1, \quad c_2, \quad c_3, \quad \ldots\ldots, \quad c_m \}$$

$$\mathbf{X}^* = \{ \ c_1, \quad c_1, \quad c_2, \quad \ldots\ldots, \quad c_9 \}$$

$$\mathbf{X}^* = \{X_1^a, X_1^b, X_1^c, X_1^d, X_1^a, X_1^b, X_1^c, X_1^d, X_2^a, X_2^b, X_2^c, X_2^d, \ldots, X_9^a, X_9^b, X_9^c, X_9^d\}$$

Confidence Intervals



Conclusions

- Many methods for computing confidence intervals underestimat the intervals
- Parametric methods underestimate the variance when data are dependent
- This is also true for the traditional Bootstrap
- For any biometrics, samples of the same person are *not* independent
- Bootstrap not so sensitive to "identical" assumption
- Introduced the ***Subsets Bootstrap***: Divide the scores up into subsets that are independent
  - Sample the set of subsets with replacement to form bootstrap sets
  - Gives realistic estimates of confidence intervals

IEEE AutoID 02, March 14-15, Westchester Mariott, Tarrytown, NY 10591

http://www.research.ibm.com/people/r/ratha/autoid.html